CARDIOVASCULAR RISK PREDICTION

Mujtaba Ali & Prateek Sachdeva
Data science trainees,
AlmaBetter, Bangalore

Abstract:

- Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels. CVD includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack). Other CVDs include stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, abnormal heart rhythms, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, peripheral artery disease, thromboembolic disease, and venous thrombosis.

1. Problem Statement

- The dataset is from an ongoing cadiovascular study on residents of the town of Framingham, Masschusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients information. It includes over 4000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

1. **Sex:** male(0) or female(1);(Nominal)
2. **Age:** age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous) education
3. **CurrentSmoker:** whether or not the patient is a current smoker (Nominal)
4. **CigsPerDay:** the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
5. **BPMeds:** whether or not the patient was on blood pressure medication (Nominal)
6. **PrevalentStroke:** whether or not the patient had previously had a stroke (Nominal)
7. **prevalentHyp:** whether or not the patient was hypertensive (Nominal)
8. **diabetes:** whether or not the patient had diabetes (Nominal)
9. **totChol:** total cholesterol level (Continuous)
10. **sysBP:** systolic blood pressure (Continuous)
11. **diaBP**: diastolic blood pressure (Continuous)
12. **BMI:** Body Mass Index (Continuous)

13. **HeartRate:** heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values).
14. **glucose:** glucose level (Continuous)
15. **10 year risk of coronary heart disease CHD** (binary: "1", means "Yes", "0" means "No") - Target Variable.

2. Introduction

By using this project, we can predict the Cardiovascular risk based on the provided features. This method can be very useful to predict the risk of cardiovascular diseases based on certain features.

3. Steps involved:

- **Exploratory Data Analysis:**
  After loading the dataset we performed this method by comparing our target variable that is cardiovascular risk with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

- **Null values Treatment:**
  Our dataset contains a fair number of null values which might tend to disturb our accuracy hence we imputed them using knn-imputer.

- **Encoding of categorical columns:**
  We used One Hot Encoding to handle nominal categorical variables. For ordinal variable, we used ordinal encoding.

- **Feature Construction:**
  We made a new feature i.e. AvgBP by taking the average of sysBP(systolic blood pressure) and diaBP(diastolic blood pressure) since they both were highly correlated features.

- **Feature Selection:**
  We checked multicollinearity using heatmap, and dropped the features having high correlation coefficient.

- **Standardization of features:**
  We used Standard-scaler to scale the distribution, which was earlier right skewed to the normal. For Outliers treatment we used robust-scaler.

- **Balancing of Dataset:**
  The dataset was highly imbalanced initially. This could have impacted the prediction results badly. So we balanced it using SMOTE (Synthetic Minority Oversampling Technique
).

- **Fitting different models:**
  For modeling we tried various classification algorithms like:

1. **Logistic Regression**
2. **K-nearest-neighbors.**
3. **Support Vector Classifier**
4. **Naïve Bayes Classifier**
5. **Decision Tree Classifier**
6. **Random Forest Classifier**
7. **XGboost Classifier.**

- **Tuning the hyperparameters for better accuracy:**
  Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting.

- **SHAP Values for features**
  We have applied SHAP value plots on the Xgboost model to determine the features that were most important while model building and the features that didn't put much weight on the performance of our model.

4. Algorithms:

- **Linear Regression:**
  By using logistic regression, we were able to obtain 67% accuracy.

- **K-neighbors Classifier:**
  We are able to obtain 84% accuracy by using KNN-classifier. We observed that accuracy improved much.

- **Decision Tree Regressor:**
  By using decision tree regressor, we got 83% accuracy.

- **Random forest:**
  We were able to obtained 89% accuracy using random forest.

- **XGBoost Regressor:**
  We were able to obtained 85% accuracy using xgboost regressor.

5. Performance metrics

- **Accuracy:**
  Accuracy will require two inputs (i) actual class labels (ii)predicted class labels. To get the class labels from probabilities(these probabilities will be probabilities of getting a HIT), you can take a threshold of 0.5. Any probability above 0.5 will be labeled as class 1 and anything less than 0.5 will be labeled as class 0.

- **Precision:**

  Precision for a label is defined as the number of true positives divided by the number of predicted positives. Report precision in percentages.

- **Recall:**

  Recall for a label is defined as the number of true positives divided by the total number of actual positives. Report recall in percentages.

- **F1-Score:**

  This is defined as the harmonic mean of precision and recall.

- **ROC_AUC Score:**

  The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

6. Hyper parameter tuning:

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

- **Grid Search CV-**Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model's performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

7. Conclusion:

That's it! We reached the end of our exercise.
Starting with loading the data so far we have done EDA, null values treatment, encoding of categorical columns, feature selection, balancing the dataset(using SMOTE), and then model building.
In all of these models our accuracy revolves in the range of 57% to 89%.
So the accuracy of our best model is 89%s(using Random Forest) which can be said to be good for this problem.